一种面向基因选择的结合 Relief-F 和决策树的 APSO 算法 *

叶超超，潘巨龙†

(中国计量大学 信息工程学院, 杭州 310018)

摘　要：由于基因表达数据高维度、高噪声、小样本的特点，基因选择一直是肿瘤分类的一大挑战。为了提高肿瘤分类的精度，同时保证基因选择的效率，提出一种结合 Relief-F 和 CART 决策树的自适应粒子群优化(APSO)算法(R-C-APSO)。该方法首先利用 Relief-F 快速过滤大量无关基因和噪声，缩小基因选择范围；然后以 CART 决策树为适应度函数，用 APSO 算法对基因进行最终搜索。通过 6 个数据集的分析实验，实验结果表明，R-C-APSO 拥有较高的分类精度和较快的基因选择速度，且具有良好的稳定性。

关键词：基因选择；自适应粒子群；决策树；肿瘤分类

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2017.08.0872

New APSO algorithm for gene selection combined with Relief-F and decision tree

Ye Chaochao, Pan Julong†

(*College of Information Engineering*, *China Jiliang University*, *Hangzhou* 310018, *China*)

Abstract: Due to the characteristics of gene expression data in high dimension, high noise and small sample, gene selection had been one of the major challenges in tumor classification. In order to improve the accuracy of tumor classification and ensure the efficiency of gene selection, this paper developed a new algorithm R-C-APSO using an adaptive particle swarm optimization (APSO) combined with Relief-F and CART decision tree. The algorithm utilized Relief-F to quickly filter a large number of irrelevant genes and noises, narrowing the gene selection range. The R-C-APSO used the CART decision tree as a fitness function, and APSO algorithm performed the final search of the genes. Based on the analysis of 6 data sets, the experimental results show that R-C-APSO can obtain high classification accuracy and fast gene selection speed, and has good stability.

Key Words: gene selection; adaptive particle swarm; decision tree; tumor classification

0　引言

DNA 微阵列技术的快速发展使得大规模基因表达数据的测量能够在单个实验中完成。基因表达数据具有维度高、噪声大、样本少的特点[1]。成千上万维的基因中对肿瘤诊断起作用的只占一小部分，如果从理论的角度直接将所有的特征基因全部用于肿瘤分类，将使得运算变得复杂且低效。所以在肿瘤分类中，基因选择是数据预处理时至关重要的一步，不但能够去除无关基因来提高分类精度，还能减少算法的计算代价[2]。

基因选择是分析微阵列表达数据的最重要任务之一，因为它可以帮助发现疾病机制，降低临床诊断成本并提高肿瘤分类精度[3]。过滤式(filter)作为特征选择方法之一，因其实现简单、计算快速而被广泛采用。过滤式特征选择通常采用统计学方法来挑选特征，该方法与分类模型无关，常用的有：互信息、t 统计量、信息增益等。文献[4]直接使用两样本 t 统计量(two sample t-statistic)来选择基因，但该方法却只能用于二分类问题;文献[5]提出了一种基于互信息和遗传算法的两阶段特征选择方法;文献[6]提出了基于信息增益和改进简化的粒子群基因选择方法。然而这些方法在处理连续型特征值时，常采用离散化的方式，这可能会导致重要特征信息的丢失;文献[7]将 Relief 引入 SNP 全基因组关联分析中，剔除了大量无关基因并获得了较高的分类精度，但 Relief 只能用于二类问题。1994 年，Kononenko 将其进行扩展，得到了 Relief-F[8]。Relief-F 不仅能处理多类问题，而且适用于特征值为连续的类型，是一种高效的数据预处理算法。

包裹式(wrapper)是另一重要的特征选择方法，包括学习器和搜索算法两个部分。著名的搜索算法有遗传算法、粒子群优化(particle swarm optimization，PSO)算法[9]等。粒子群优化算法由 James Kennedy 和 Russel Eberhart 于 1995 年提出，源于对鸟群捕食行为的研究，因其参数少、实现简单且具有较好的全局搜索能力而被广泛应用[10-13]。文献[14]提出一种新的局部搜索策略，并使用粒子群优化算法来选择基因子集。文献[15]将基本

基金项目：国家自然科学基金资助项目（61272315）

作者简介：叶超超（1994-），男，浙江余姚人，硕士研究生，主要研究方向为机器学习、深度学习（1215172385@qq.com）；潘巨龙（1965-），男（通信作者），教授，博士，主要研究方向为无线传感网络、移动计算和智能计算（pjl@cjlu.edu.cn）.

粒子群算法与决策树分类器相结合来对肿瘤进行分类;文献[16]提出一种将支持向量机(support vector machine，SVM)和改进二元粒子群优化算法相结合的方法来选择近最优特征子集;文献[17]通过将相关系数与粒子群优化算法相结合，提出一种混合特征选择方法，在三个肿瘤数据集上获得了较高的分类精度。上述研究都表明 PSO 是一种优秀的基因选择算法。然而当特征维数很高、搜索空间中拥有多个局部极值时，PSO 算法就会出现过早收敛、陷入局部最优和计算耗时等问题。通过上面的分析，本文提出一种新型高效的基因选择算法 R-C-APSO。一方面，利用 Relief-F 过滤快速的特点来对基因进行初选，限制基因数量以加快后期搜索算法的速度。另一方面，在 PSO 的基础上引入自适应惯性权重，使用全局搜索能力更强的自适应粒子群优化(adaptive particle swarm optimization，APSO)算法来对基因进行最终的挑选。同时，本文选用分类与回归树(classification and regression tree，CART)作为 APSO 的适应度函数。将 CART 与 APSO 相结合的算法目前研究成果还不多。CART 决策树因其本身具有特征选择的功能，故将其作为 APSO 的适应度函数来最终验证基因选择的有效性，将进一步提高肿瘤的分类精度。

1 相关概念

1.1 Relief-F

Relief-F 是一种运行效率很高的特征选择算法，它在 Relief 算法的基础上扩展了多类别数据处理的功能，同时解决了数据存在噪声和不完整的问题[8]。该算法通过计算每个特征上的一个“相关统计量”来度量特征的重要性。

假设有样本集合 $S=\{(\boldsymbol{a}_1,y_1),(\boldsymbol{a}_2,y_2),\ldots,(\boldsymbol{a}_m,y_m)\}$，共 L 个类别。从集合中随机选择一个样本 $\boldsymbol{a}_i$，假定它的类别为 k($k\in\{1,2,...,L\}$)。首先在 $\boldsymbol{a}_i$ 同类别样本中找到与之最近的样本 $\boldsymbol{a}_{i,nh}$，将其称为猜中近邻；然后再找出其他每个类别中与 $\boldsymbol{a}_i$ 最近的样本 $\boldsymbol{a}_{i,l,nm}$ ($l=1,2,...,L;l\neq k$)，将其称为猜错近邻。那么，特征 j 的一次相关统计量可定义为

$$\varphi^j=-\left(a_i^j-a_{i,nh}^j\right)^2+\sum_{l\neq k}\left(p_l\times\left(a_i^j-a_{i,l,nm}^j\right)^2\right) \tag{1}$$

p_l 表示类别为 l 的样本在集合 S 中的比例。根据(1)式，计算本次各个特征的相关统计量。重复上述操作抽样 M 次，将每次计算得到的结果对应相加即可得到每个特征的相关统计量。一个特征的相关统计量越大，表示这个特征在分类中越重要。通过对所有特征进行排序，然后设定阈值或特征选择个数，便可得到过滤后的特征子集。

1.2 粒子群优化(PSO)算法

粒子群优化(PSO)算法是一种启发式的全局优化方法，在多维空间中通过模拟鸟群的社会行为来实现一个近似最优解[18]。PSO 算法中的每一个对象被称作粒子，每个粒子具有位置和速度两个属性。假定算法运行在 n 维空间中，则粒子 i 的第 p 次迭代位置和速度可以分别表示为 $\boldsymbol{x}_i^p=[x_{i1}^p,x_{i2}^p,\ldots,x_{in}^p]$、$\boldsymbol{v}_i^p=\left[v_{i1}^p,v_{i2}^p,\ldots,v_{in}^p\right]$。在迭代过程中，每个粒子根据适应度函数的计算值来判断当前位置的优劣，用 $pbest_{id}^p$ 表示粒子 i 在前 p 次迭代过程中的最佳位置在第 d 维上的值，用 $gbest_d^p$ 表示所有粒子在前 p 次迭代过程中的最佳位置在第 d 维上的值。粒子通过当前位置、速度和上述两个最佳位置来决定下一步的位置和速度，粒子 i 的更新公式如下：

$$v_{id}^{p+1}=\omega v_{id}^p+c_1r_1\left(pbest_{id}^p-x_{id}^p\right)+c_2r_2\left(gbest_d^p-x_{id}^p\right) \tag{2}$$

$$x_{id}^{p+1}=\begin{cases}1,\mathrm{sigmoid}\left(v_{id}^{p+1}\right)=1\Big/\left(1+e^{-v_{id}^{p+1}}\right)>\mathrm{U}(0,1)\\0,\mathrm{otherwise}\end{cases} \tag{3}$$

其中: c_1，c_2 称为学习系数；r_1 和 r_2 取 0 到 1 之间的随机数；ω 是粒子迭代时的惯性权重[19]，一般取常数；$\mathrm{U}(0,1)$ 表示 0 到 1 之间服从均匀分布的随机数。另外，迭代过程中每个粒子的速度必须限制在 $[\boldsymbol{V}_{\min},\boldsymbol{V}_{\max}]$ 之间[9,15]。

2 结合 Relief-F 和决策树的 APSO 算法

2.1 自适应粒子群优化(APSO)算法

PSO 算法虽然拥有较强的搜索能力，但在面对多极值问题时，常因陷入局部最优而错过最优解[13]。为了进一步增强粒子的整体搜索能力，本文引入自适应惯性权重 ω^p 来更新粒子的速度：

$$v_{id}^{p+1}=\omega^p v_{id}^p+c_1r_1\left(pbest_{id}^p-x_{id}^p\right)+c_2r_2\left(gbest_d^p-x_{id}^p\right) \tag{4}$$

$$\omega^p=\left(\omega_{\max}-\omega_{\min}\right)\times\exp\left(-\left(\eta\times p/P_{\max}\right)^2\right)+\omega_{\min} \tag{5}$$

其中: $\omega_{\max}$ 和 $\omega_{\min}$ 分别为最大和最小惯性权重；η 为[20,55]的经验值；p 为当前迭代次数，$P_{\max}$ 为预设的最大迭代次数。由(5)式可以看出，ω^p 的值随着迭代的进行从 $\omega_{\max}$ 最终降到 $\omega_{\min}$。惯性权重具有平衡粒子群算法的全局搜索和局部搜索功能[13]。当 ω^p 较大时，可以提高 APSO 的全局搜索能力；当 ω^p 较小时，APSO 拥有较强的局部搜索能力。

2.2 适应度函数

适应度函数是评价粒子个体质量的指标，适应度函数选取的好坏对 APSO 算法的收敛速度及能否找到最优解具有很大的影响。本文将 CART 决策树作为适应度函数，因其具有特征选择的能力，这将使特征子集得到更好的验证，粒子个体质量得到更好的评价。

假设存在一个含 n 类样本的样本集 D，第 m 类样本的比例为 p_m ($m=1,2,\ldots,n$),那么 D 的纯度可以用下式衡量：

$$\mathrm{Gini}(D)=\sum_{m=1}^{n}\sum_{m'\neq m}p_m p_{m'}=1-\sum_{m=1}^{n}p_m^2 \tag{6}$$

D 的基尼值 $\mathrm{Gini}(D)$ 与其纯度呈负相关关系，其值越小就意味着 D 的纯度越高。

CART 决策树在选择划分属性时，使用的指标为基尼指数[20]。假定用属性 b 对样本集 D 进行划分并产生 F 个分支，将分支 f 的样本集记为 D^f，则属性 b 的基尼指数定义为

$$\text{GiniIndex}(D,b)=\sum_{f=1}^{F}\frac{|D^f|}{|D|}\text{Gini}(D^f) \tag{7}$$

$|D|$ 和 $|D^f|$ 分别表示集合 D 和 D^f 的样本个数。CART 决策树选择基尼指数最小的属性作为划分属性。属性的基尼指数越小，则划分后各个分支集合的样本纯度越高。

2.3　算法描述

为了提高肿瘤分类的精度同时保证基因选择的效率，本文在 PSO 算法的基础上提出了一种新的基因选择算法：一种面向基因选择的结合 Relief-F 和决策树的 APSO 算法(R-C-APSO)。该方法充分利用了 Relief-F 的过滤性能、自适应粒子群优化(APSO)算法的非线性搜索能力和 CART 决策树本身的特征选择优势。首先，算法使用 Relief-F 根据(1)式计算每个基因的相关统计量，按计算结果从大到小对基因进行排序。一个基因的相关统计量越大，那么这个基因对肿瘤分类起到的作用就越大。然后，对排序后的基因进行筛选，只保留排序靠前的 T 个基因，去除其余大量无关基因和噪声。最后，使用上一步筛选出来的基因对 APSO 的种群进行初始化，利用搜索能力较强的 APSO 来选择最终的基因子集。由于前期 Relief-F 大大降低了基因的维度，这将为 APSO 的搜索节省大量时间。

$$v_{id}^0=\text{U}(0,1) \tag{8}$$

$$x_{id}^0=\begin{cases}1, \text{U}(0,1)>0.5\\ 0, \text{otherwise}\end{cases} \tag{9}$$

在 APSO 的种群初始化阶段，粒子 i 的初始速度和初始位置在第 d 维上的值通过通过式(8)和(9)计算获取，其中 $\text{U}(0,1)$ 表示 0 到 1 之间服从均匀分布的随机数。另外，在 APSO 搜索阶段，将分类器 CART 决策树作为适应度函数，用其分类精度来对每次搜索到的基因子集进行评价。CART 决策树具有特征选择的功能，且模型构建速度较快，作为适应度函数来评价基因子集将进一步提升肿瘤的分类精度和基因选择的效率。

R-C-APSO 算法具体实现过程描述如下：

a)对每个样本的原始特征进行 0 均值标准化处理；

b)使用 Relief-F 算法根据(1)式计算各个基因的相关统计量大小，其中,重复抽样次数 M 为样本个数，平均每个样本被抽到一次；

c)根据相关统计量大小对所有基因进行排序，只保留前 T 个基因，这里 T 取 100；

d)设置 APSO 的种群个数 N 和最大迭代次数 P_{max}，并对种群进行初始化。这里 N 取 20，P_{max} 取 100；

e)设置 APSO 的学习系数 c_1，c_2 均为 2；粒子最大速度 V_{max} 为 4，最小速度 V_{min} 为-4；最大惯性权重 ω_{max} 和最小惯性权重 ω_{min} 参考文献[21]，分别取 0.9 和 0.4；

f)根据式(5)计算出自适应惯性权重，然后用分类器 CART 决策树计算各个粒子的适应度函数值；

g)根据所有粒子的适应度函数值更新种群的最佳位置和每个粒子的最佳位置，并根据式(4)、(3)更新各个粒子的速度和位置；

h)判断当前是否达到最大迭代次数，若没有达到，则转至步骤 f)继续迭代；否则，退出结束。

3　实验结果及分析

为了验证 R-C-APSO 的有效性，本文选用 6 个公开的基因数据集来进行实验，分别是 SRBCT、Brain_Tumor1、9_Tumors、Prostate_Tumor、Brain_Tumor2、DLBCL。各基因数据集的详情如表 1 所示。此次实验在 64 位 windows7 操作系统上完成，代码运行环境为 python2.7, 计算机处理器为 i5-3230M@2.6 GHz，内存 8 GB。

实验中，以 CART 决策树作为分类器，对每一个基因数据集进行 10 次实验。每次实验采用五折交叉验证，以 10 次实验结果的平均值作为最终的分类精度。为了更好地评价 R-C-APSO 的性能，本文还和如下算法进行了对比实验，包括：CART 算法；Relief-F 算法；以 CART 为适应度函数的 PSO 算法(CART-PSO)；SVM 算法。

表 2 给出了 6 个数据集在 5 种算法下的实验结果。

表 1　数据集描述

数据集	特征维度	样本总数	类别总数
SRBCT	2038	83	4
Brain_Tumor1	5920	90	5
9_Tumors	5726	60	9
Prostate_Tumor	10509	102	2
Brain_Tumor2	10367	50	4
DLBCL	5469	77	2

表 2　R-C-APSO 与对比算法在各数据集上的平均分类精度　　/%

算法	SRBCT	Brain_Tumor1	9_Tumors	Prostate_Tumor	Brain_Tumor2	DLBCL
CART	74.72	69.65	33.08	83.04	57.85	79.10
Relief-F	80.56	75.20	34.98	86.27	63.33	86.28
CART-PSO	93.29	82.51	54.13	93.53	77.50	94.72
SVM	89.78	79.64	51.39	88.51	76.18	95.14
R-C-APSO	**93.74**	**85.79**	**57.01**	**94.30**	**84.15**	**96.70**

从表中可以看到，本文提出的 R-C-APSO 算法在各个数据集上的分类精度是最高的。CART-PSO 与经典的 SVM 相比，除了在 Brain_Tumors 和 DLBCL 数据集上两者的分类精度接近，在其他数据集上 CART-PSO 的表现均优于 SVM。这说明 CART 决策树与 PSO 相结合已是一种较好的基因选择方法。另外，从表 2 的 2、3 两行可以看出，经过 Relief-F 操作后的 CART 算法在各数据集上的分类精度有了明显的提升，这从侧面说明了 Relief-F 适用于基因初选，能过滤掉大量的无关基因和噪声。

CART-PSO 和 R-C-APSO 是上述 5 种算法中分类精度最高的两种，且都是基于粒子群优化算法。为了进一步比较这两种算法的优劣，本文在实验时还记录了两种算法的 CPU 运行时间和迭代过程，如表 3 和图 1~3 所示。

表 3 CART-PSO 与 R-C-APSO 在各数据集上的 CPU 运行时间(秒)

数据集	CART-PSO	R-C-APSO
SRBCT	75.89	6.19
Brain_Tumor1	243.00	9.83
9_Tumors	203.88	6.63
Prostate_Tumor	310.34	10.80
Brain_Tumor2	188.29	6.32
DLBCL	113.08	6.86

从表3可以直观地看到，R-C-APSO在各个数据集上的CPU运行时间均远远小于 CART-PSO。在自适应粒子群优化算法中，适应度函数的复杂度是该算法复杂度的主要组成部分。本文以 CART 决策树作为适应度函数，其复杂度随着特征维度的升高而增大。因此，当采用 Relief-F 对基因进行降维后，APSO 的复杂度便得到了有效降低，这大大提高了 R-C-APSO 的时间效率。

此外，从图 1~3 可以清楚地发现，与 CART-PSO 相比，R-C-APSO 在 6 个数据集上均拥有更好的搜索起点。且在之后的迭代过程中，R-C-APSO 的分类精度也始终大于 CART-PSO。所以，R-C-APSO 在分类性能上具有较强的稳定性。同时本文还可以看到，两种算法在获取相同分类精度的条件下，R-C-APSO 只需要更少的时间和迭代次数。这在计算资源有限或时间要求较高的情况下，R-C-APSO 具有更大的优势。

综上分析，本文提出的 R-C-APSO 算法在分类精度上都要高于其他四种算法。与使用粒子群优化算法的 CART-PSO 相比，R-C-APSO 还具有基因选择速度快，稳定性好的优点。

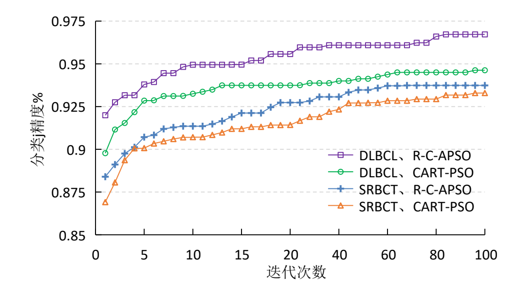


图 1 SRBCT、DLBCL 数据集上两种算法的迭代过程

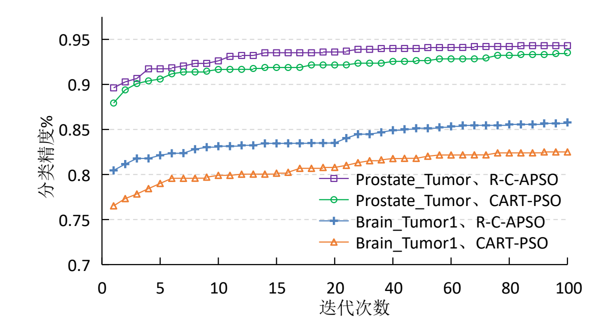


图 2 Brain_Tumor1、Prostate_Tumor 数据集上两种算法的迭代过程

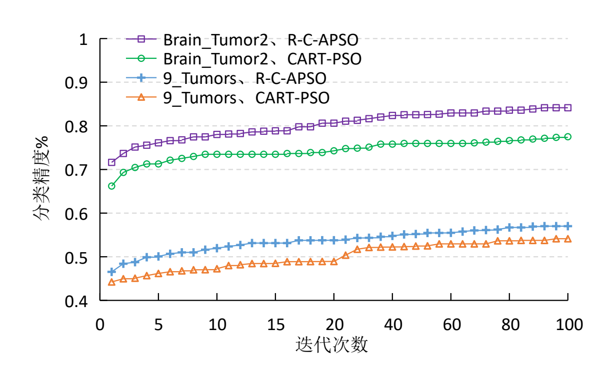


图 3 9_Tumors、Brain_Tumor2 数据集上两种算法的迭代过程

4 结束语

本文充分利用自适应粒子群优化算法的非线性搜索能力、Relief-F 的过滤性能和 CART 决策树的分割优势，提出了一种面向基因选择的结合 Relief-F 和决策树的 APSO 算法。该算法不但提高了肿瘤分类的精度，而且在运算时间上远小于粒子群优化算法，具有较好的稳定性。通过实验分析表明，本文提出的 R-C-APSO 是一种有效的基因选择算法。

本文在自适应粒子群优化算法的部分参数选择上，如最大惯性权重、最小惯性权重等主要参考了以往的文献，接下来将进行更多的研究和实验，通过优化这些参数来提高算法的收敛速度和全局搜索能力。另外，在基因选择方法的评价方面，本文主要从最终肿瘤分类的效果上来进行，对所选择基因的个数、特点未进行深入的分析，这将是本文接下来的工作。

参考文献：

[1] Ghosh S, Mitra S, Dattagupta R. Fuzzy clustering with biological knowledge for gene selection [J]. Applied Soft Computing, 2014, 16: 102-111.

[2] Li S, Wu X, Tan M. Gene selection using hybrid particle swarm optimization and genetic algorithm [J]. Soft Computing, 2008, 12 (11): 1039-1048.

[3] Chen Y, Wang L, Li L, et al. Informative gene selection and the direct classification of tumors based on relative simplicity [J]. BmcBioinformatics, 2016, 17 (1): 1-16.

[4] Varma S, Simon R. Iterative class discovery and feature selection using Minimal Spanning Trees [J]. Bmc Bioinformatics, 2004, 5 (1): 126-126.

[5] 裘国永，王娜，汪万紫. 基于互信息和遗传算法的两阶段特征选择方法

[J]. 计算机应用研究, 2012, 29 (8): 2903-2905.

[6] Lai C M, Yeh W C, Chang C Y. Gene selection using information gain and improved simplified swarm optimization [J]. Neurocomputing, 2016, 218: 331-338.

[7] 吴红霞，吴悦，刘宗田，等. 基于 Relief 和 SVM-RFE 的组合式 SNP 特征选择 [J]. 计算机应用研究, 2012, 29 (6): 2074-2077.

[8] Kononenko I. Estimating attributes: analysis and extensions of RELIEF [C]// Proc of European Conference on Machine Learning on Machine Learning. New York: Springer-Verlag, 1994: 171-182.

[9] Zhang Y, Wang S, Ji G. A comprehensive survey on particle swarm optimization algorithm and its applications [J]. MathematicalProblems in Engineering, 2015, 2015 (1): 1-38.

[10] Ghamisi P, Benediktsson J A. Feature selection based on hybridization of genetic algorithm and particle swarm optimization [J]. IEEE Geoscience & Remote Sensing Letters, 2014, 12 (2): 309-313.

[11] Xi M, Sun J, Li L, et al. Cancer feature selection and classification using a binary quantum-behaved particle swarm optimization and support vector machine [J]. Computational & Mathematical Methods in Medicine, 2016, 2016 (9): 1-9.

[12] Teng X, Dong H, Zhou X. Adaptive feature selection using v-shaped binary particle swarm optimization [J]. Plos One, 2017, 12 (3) .

[13] Lu H, Du B, Liu J, et al. A kernel extreme learning machine algorithm based on improved particle swam optimization [J]. Memetic Computing, 2017, 9 (2): 121-128.

[14] Moradi P, Gholampour M. A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy [J]. Applied Soft Computing, 2016, 43: 117-130.

[15] Chen K H, Wang K J, Tsai M L, et al. Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm [J]. BMC Bioinformatics, 2014, 15 (1): 49.

[16] Mohamad M S, Omatu S, Deris S, et al. Particle swarm optimization for gene selection in classifying cancer classes [J]. Artificial Life & Robotics, 2009, 14 (1): 16-19.

[17] Chinnaswamy A, Srinivasan R. Hybrid feature selection using correlation coefficient and particle swarm optimization on microarray gene expression data [M]// Innovations in Bio-Inspired Computing and Applications. [S. l.] : Springer International Publishing, 2016: 1117-1124.

[18] Sheikhpour R, Sarram M A, Sheikhpour R. Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer [J]. Applied Soft Computing, 2016, 40: 113-131.

[19] Taherkhani M, Safabakhsh R. A novel stability-based adaptive inertia weight for particle swarm optimization [J]. Applied Soft Computing, 2016, 38: 281-295.

[20] Manek A S, Shenoy P D, Mohan M C, et al. Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier [J]. World Wide Web-internet & Web Information Systems, 2017, 20 (2): 135-154.

[21] Shi Y, Eberhart R C. Parameter selection in particle swarm optimization [C]// International Conference on Evolutionary Programming Vii. New York: Springer-Verlag, 1998: 591-600.